计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2026,35(2):262−268 [doi: 10.15888/j.cnki.csa.010076] [CSTR: 32024.14.csa.010076]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

智慧城市边缘场景下的人-物一体化检测①

白　珊[1], 单卓然[2]

[1](武汉光谷职业学院 人居工程设计学院, 武汉 430200)
[2](华中科技大学 建筑与城市规划学院, 武汉 430074)
通信作者: 单卓然, E-mail: 371760860@qq.com

摘　要: 随着智慧城市建设的深入推进, 建筑边缘区域的安全问题日益严峻, 人员意外坠落与高空坠物事件频发, 亟需更加智能、高效的监测手段. 针对当前目标检测方法在小目标、遮挡目标及高速运动目标识别中的时序建模能力不足的问题, 本文提出一种融合多种时间语义增强机制的视频检测框架, 用于实现人员与坠落物的一体化检测. 所提方法在 Faster R-CNN 主干结构上集成了 3 种时序感知模块: 运动感知模块 (MAM)、时间区域兴趣点对齐操作符 (TROI Align) 和序列级语义聚合头部 (SELSA Head), 分别从运动显著性建模、空间对齐和语义聚合这 3 个角度, 提升模型对复杂时序场景中动态目标的感知能力. 为支撑模型训练与评估, 本文构建了一个覆盖建筑边缘多场景、多类风险目标的视频数据集. 实验结果表明, 本文方法在“人员临边行为检测”与“高空坠物检测”两个子任务中表现出良好效果, 展现出良好的跨任务鲁棒性与实际应用潜力.

关键词: 智慧城市; 边缘检测; 人员检测; 坠落物检测; 深度学习

引用格式: 白珊,单卓然.智慧城市边缘场景下的人-物一体化检测.计算机系统应用,2026,35(2):262–268. http://www.c-s-a.org.cn/1003-3254/10076.html

Person-object Integrated Detection in Smart City Edge Scenarios

BAI Shan[1], SHAN Zhuo-Ran[2]

[1](School of Human Settlements and Engineering Design, Wuhan Guanggu Vocational College, Wuhan 430200, China)
[2](School of Architecture and Urban Planning, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: With the continuous advancement of smart city development, safety issues in building edge areas have become increasingly severe, as incidents of accidental falls and falling objects occur frequently. There is an urgent need for more intelligent and efficient monitoring solutions. To address the limited temporal modeling capabilities of current object detection methods, particularly in recognizing small, occluded, and fast-moving targets, this study proposes a video detection framework that integrates multiple temporal semantic enhancement mechanisms for the unified detection of both people and falling objects. The proposed method is built upon a faster R-CNN backbone and incorporates three temporal-aware modules: motion-aware module (MAM), temporal region of interest align (TROI Align), and sequence-level semantic aggregation head (SELSA Head). These modules enhance the model’s perception of dynamic objects in complex temporal scenarios from three perspectives: motion saliency modeling, spatial alignment, and semantic aggregation. To support model training and evaluation, a dedicated video dataset covering multiple building edge scenarios and various types of risk targets is constructed. Experimental results demonstrate that the proposed method achieves strong performance in both “detection of personnel behavior at building edges” and “falling object detection” tasks, showing excellent cross-task robustness and practical application potential.

Key words: smart city; edge detection; personnel detection; falling object detection; deep learning

① 收稿时间: 2025-07-25; 修改时间: 2025-08-28; 采用时间: 2025-09-15; csa 在线出版时间: 2025-12-26
CNKI 网络首发时间: 2025-12-29

1 引言

1.1 智慧城市背景与边缘安全挑战

智慧城市是新一代信息技术与城市运行体系深度融合的产物, 正推动城市管理迈向数字化、智能化的新阶段[1,2]. 随着城市空间的垂直扩展, 高层建筑数量激增, 阳台、天台、窗台等建筑边缘空间逐渐演变为潜在的高风险区域. 近年来, 因人员临边行为导致的意外坠落, 以及高空抛掷、坠落物体引发的伤害事件屡有发生, 不仅严重威胁居民生命财产安全, 也对公共安全管理提出了更高要求.

边缘安全风险主要来源于两类典型场景: 一是人员在建筑边缘区域的高危行为, 如攀爬、探身、倚靠护栏等, 存在坠落风险; 二是物体从高空坠落或被抛掷, 包括瓶子、书本、工具、包装物等, 极易对过往行人造成突发性伤害. 这类事件具有目标体积小、运动速度快、遮挡频繁、时间持续短等特点, 导致传统静态图像检测方法或人工监控方式难以胜任[3].

尽管城市中已广泛部署视频监控系统, 但在实际应用中仍面临诸多挑战. 一方面, 目标在视频中往往以小尺度、姿态变化大、遮挡严重的形式出现; 另一方面, 复杂光照、动态背景与运动模糊等因素也会影响检测鲁棒性. 在此背景下, 构建具备时序信息建模能力的智能检测框架, 以实现对建筑边缘风险事件的高效识别与实时预警, 成为智慧城市安全监控中的关键课题[4].

1.2 研究现状与现有数据集局限

近年来, 随着视频目标检测与行为识别技术的发展, 计算机视觉在城市安全领域的应用逐渐深入, 尤其在人群密集、交通复杂和监控盲区的智能感知方面取得了一定成果. 在边缘安全相关领域, 研究主要集中于通用目标检测、小目标检测、运动目标检测与视频行为识别等方向.

一方面, 基于卷积神经网络 (CNN) 的目标检测方法已取得显著进展. 经典的两阶段检测器如 Faster R-CNN[5]以及一阶段检测器如 YOLO 系列[6]和 RetinaNet[7]等, 在多个公开图像数据集 (如 MS COCO[8]、PASCAL VOC[9]) 上均取得良好表现. 然而, 这些方法多依赖静态图像, 对于复杂视频环境中的遮挡、模糊与多目标干扰的鲁棒性仍有限.

另一方面, 为提升视频中的目标检测精度, 部分研究开始引入时间建模机制, 如 FGFA[10]、MEGA[11]和 TROI[12]等方法通过帧间对齐或语义聚合, 有效增强了检测对遮挡与小目标的适应能力. 同时, 运动目标检测 (MOD) 领域也积累了较多成果, 但大多集中在前景提取或背景建模上, 面向高层建筑边缘或高空环境的安全监控任务较少.

在数据资源方面, 目前广泛使用的检测数据集多为通用场景设计, 如 COCO[13]、CityPersons[14]和 CrowdHuman[15], 在目标尺度、场景布局与视频属性上与真实城市边缘场景存在较大差异. 特别是在建筑边缘区域, 目标普遍尺度较小、姿态多样、遮挡严重, 上述数据集难以覆盖类似场景. 即使部分行人检测数据集如 Caltech[16]、KITTI[17]在城市交通场景中表现良好, 也难以迁移至高空事件检测任务.

为弥补现有方法在边缘视频场景中对时序动态理解能力的不足, 本文从方法结构设计角度出发, 整合 EBPersons[18]与 FADE[19]两类经典检测框架中的关键时序建模组件, 提出一种统一的人-物一体化视频检测方法. 所采用的 3 个模块——MAM (运动感知模块)、TROI Align (时间区域兴趣点对齐操作符) 与 SELSA Head (序列级语义聚合头部) 分别针对不同维度的时间语义进行建模, 协同提升模型在小目标、遮挡、运动模糊等复杂场景下的鲁棒性与时效性. 与传统单帧检测或局部时间建模方法不同, 本文实现了从运动动态、空间对齐到语义聚合的全链路时序建模机制整合, 适配智慧城市建筑边缘环境下人员与物体混合目标的高精度检测需求.

1.3 本文主要贡献

针对建筑边缘区域频发的多类高风险事件, 本文提出一种融合时间信息建模机制的人-物一体化检测框架, 在统一任务目标、模块选择与评估平台等方面做出以下贡献.

(1) 任务融合建模: 本文将“人员临边行为检测”与“高空坠物检测”纳入统一任务框架, 提出“人-物一体化边缘安全检测”这一问题, 系统建模城市边缘空间中不同类型高风险目标的共同检测需求, 拓展了视频目标检测在实际城市安全场景中的应用边界.

(2) 时序语义增强设计: 在 Faster R-CNN 主干结构基础上, 本文引入 3 种具备时序建模能力的关键模块——MAM (运动感知模块)、TROI Align (时间区域兴趣点对齐操作符) 与 SELSA Head (序列级语义聚合头部), 构建“运动-空间-语义”这 3 层时序感知机制, 全

面提升模型在动态复杂视频场景中的检测鲁棒性.

(3) 数据集构建与验证评估: 本文整合现有数据集, 构建了一个专用于边缘安全检测的视频数据集, 覆盖多类城市建筑外部监控场景, 统一标注人员与坠物两类目标. 在此数据集上完成多任务对比实验与模块消融分析, 验证所提方法在多个关键指标下的有效性与通用性.

2 人-物一体化检测算法

本节将详细阐述本文提出的面向智慧城市边缘场景的人-物一体化检测算法. 该算法以先进的深度学习框架为基础, 融合了时间信息处理模块, 旨在高效、鲁棒地检测高楼边缘区域的人员及坠落物.

2.1 整体网络架构

本文提出的人-物一体化检测算法基于 Faster R-CNN 框架构建, 整体结构包括主干特征提取网络、区域提议网络 (RPN) 及多级时间建模模块, 如图 1 所示.

为增强模型在小目标、遮挡与运动目标场景下的鲁棒性, 首先在主干网络阶段引入运动感知模块 (MAM), 通过运动掩码引导特征聚焦动态区域; 随后在 RPN 之后, 依次加入时间区域兴趣点对齐操作符 (TROI Align) 与序列级语义聚合头部 (SELSA Head) 模块, 分别用于帧间空间对齐与语义时序融合. 三者协同建模时间信息, 有效提升检测对复杂动态场景的适应能力. 最终, 时序增强后的特征被送入检测头部, 实现对人员行为与坠落物的统一识别与定位.

2.2 运动感知模块

为增强时序建模能力、提升对运动目标的感知效果, 本文在特征提取阶段引入运动感知模块 (motion-aware module, MAM), 通过外部运动信息引导网络关注视频中的动态区域, 从而增强特征表达的鲁棒性, 如图 2 所示.

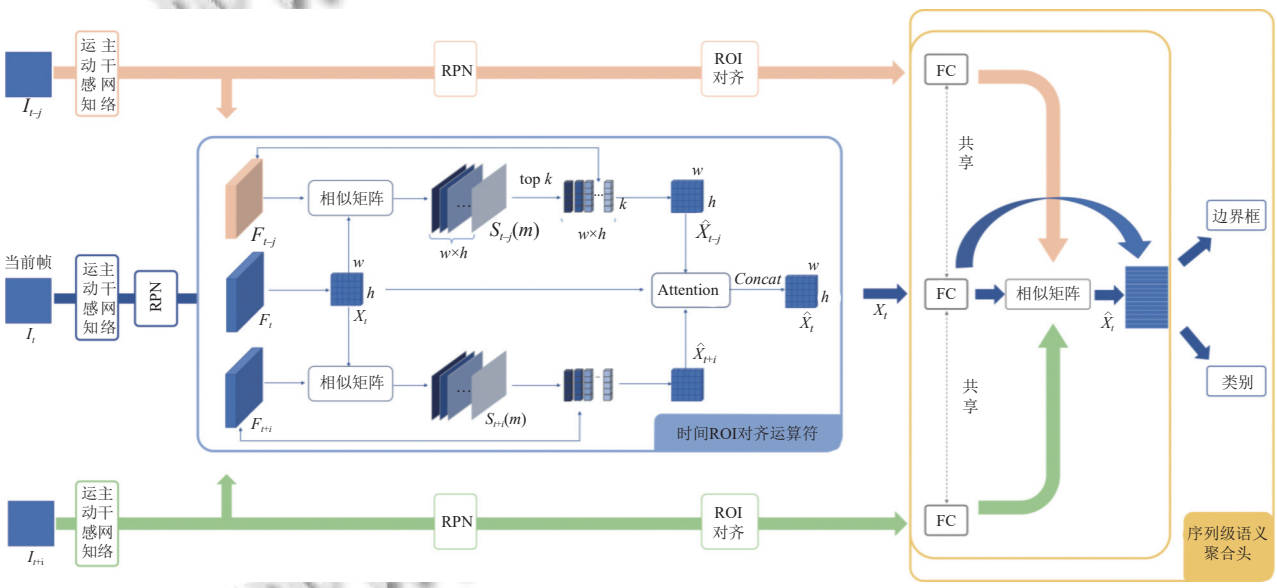

图 1 整体网络框架

该模块以帧间运动信息为基础, 通过高斯混合背景建模算法 MOG2[20]生成每帧的运动掩码$M \in \mathbb{R}^{H\times W\times 1}$, 用于标记显著运动区域. 为提高召回率, MOG2 使用较低阈值, 以覆盖弱运动区域. 随后, 运动掩码与主干网络输出的特征图$F \in \mathbb{R}^{H\times W\times C}$一同输入注意力模块. 本文对特征图执行最大池化与平均池化操作, 并将两者与运动掩码拼接后, 通过 7×7 卷积层和 Sigmoid 激活生成运动注意力图$A \in \mathbb{R}^{H\times W\times 1}$, 计算公式如式 (1):

$$A = \sigma(Conv(Concat[AvgPool(F), MaxPool(F), M])) \tag{1}$$

其中, $Conv(\cdot)$表示 7×7 的卷积操作, $\sigma(\cdot)$为 Sigmoid 函数, $Concat$ 表示通道维拼接.

随后, 将注意力图 A 与原始特征图 F 进行逐元素乘法, 生成融合运动显著性信息的特征图:

$$F^{\mathrm{MAM}} = A \odot F \tag{2}$$

其中, ⊙表示逐元素乘法操作.

MAM 可部署于主干网络的多个尺度层级, 对不同分辨率特征进行动态引导. 在与后续的 TROI Align 和 SELSA Head 模块串联时, MAM 提供了“运动优先”的

建模起点, 有效突出运动目标区域; 而后两者分别在空间位置与语义表达层面对时序进行补充建模. 三者结合, 构成本文“运动-空间-语义”三级时序建模机制, 显著提升模型对高空坠落物等快速短时目标的检测鲁棒性.

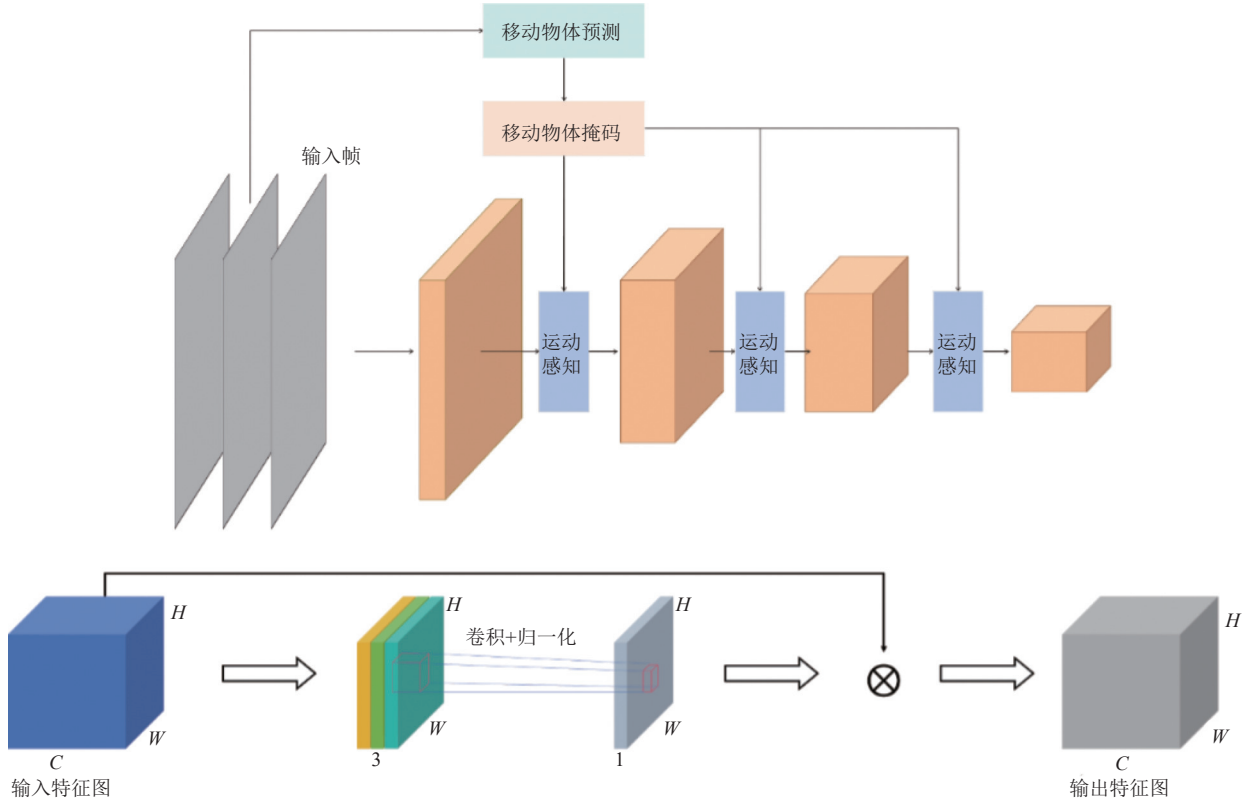

图 2 运动感知模块详细结构

2.3 时间区域兴趣点对齐操作符

在高楼边缘等复杂场景中, 目标往往存在遮挡严重、姿态变化大、光照不稳定等问题, 单帧检测难以获得稳定特征. 为此, 本文引入时间区域兴趣点对齐操作符 (TROI Align)[12], 用于在相邻帧中建立跨帧区域级对齐关系, 增强时序信息建模能力.

具体地, 设当前帧的 ROI 特征为X_t, 支持帧$t+i$特征为F_{t+i}, 首先在各帧上独立执行 ROI Align 操作. 随后, 通过点乘计算空间位置m上的相似性.

$$S_{t+i}(m)=X_t(m)\otimes[F_{t+i}]^{\mathrm{T}} \tag{3}$$

其中, $\otimes$表示矩阵乘法, m表示X_t在空间上的位置. 接着, 从支持帧F_{t+i}中选择与 $X_t(m)$具有最高相似度的K个空间位置点. 这些选定点的加权和被用作$X_t(m)$在时间$t+i$处的支持特征. 通过这种方式, 我们构建了一个增强的支持特征图$\hat{X}_{t+i}$. 接下来, 当前帧的 ROI 特征X_t及其来自不同支持帧的增强支持特征$\left\{\hat{X}_{t+i}\right\}_{i=-T/2}^{T/2}$ (其中, T 表示时间窗口长度) 被沿着通道维度分成N组. 对于第n组特征$\left\{\hat{X}_{t+i}^{(n)}\right\}_{i=-T/2}^{T/2}$, 它们通过一个注意力模块沿着时间维度进行聚合. 最终, 将各组聚合后的结果进行拼接, 得到当前帧最终的时间感知 ROI 特征X_t. 该模块能有效捕捉跨帧的区域一致性, 提升模型在检测姿态变化、遮挡目标及高速坠落物体时的稳定性与准确性.

2.4 序列级语义聚合头部

为进一步深度挖掘视频序列中的时间信息, 本文在时间区域兴趣点对齐操作符的输出之后, 引入了序列级语义聚合头部 (SELSA Head)[21]通过多帧之间的语义关联构建鲁棒特征表达, 提升检测在复杂动态场景中的稳定性.

SELSA Head 首先对当前帧t的时间感知 ROI 特征$\bar{X}_t$以及来自其他帧的原始 ROI 特征X_{t+i}进行处理. 这些特征会通过一个共享的全连接层$\phi(\cdot)$进行高维投影, 以提取其核心语义信息. 随后, 模型计算这些投影后特征之间的语义相似度W_t^{t+i}, 其计算方式如式 (4) 所示:

$$W_t^{t+i}=\phi(\bar{X}_t)\otimes[\phi(X_{t+i})]^{\mathrm{T}} \tag{4}$$

其中, $\phi(\cdot)$再次代表全连接层操作. 此语义相似度矩阵 W_t^{t+i}经由 Softmax 函数沿第 1 个维度进行归一化, 其输出的权重反映了各帧特征对当前帧语义聚合的贡献度. 最终, 这些归一化权重被用于加权聚合来自其他帧的原始特征X_{t+i}. 聚合后的特征与当前帧的投影特征$\phi(\bar{X}_t)$相结合, 形成最终的序列级聚合特征$\tilde{X}_t$, 如式 (3) 所示:

$$\tilde{X}_t = \phi(\bar{X}_t) + \sum_{i=-\frac{T}{2}}^{\frac{T}{2}} W_t^{t+i} \otimes [X_{t+i}]^{\mathrm{T}} \tag{5}$$

该聚合过程在 SELSA Head 中重复执行 3 次, 形成多层语义聚合结构, 有效捕捉视频中的长期时间上下文. 在面对如光照突变、遮挡干扰或瞬时运动等挑战时, 该模块可显著提升检测精度与鲁棒性, 降低误检率, 特别适用于高空坠物和人员行为等短时高动态场景.

3 实验分析

为验证本文提出的人-物一体化边缘安全检测方法的有效性与通用性, 本文在自建的视频数据集上开展了系统实验. 通过多种对比实验与模块消融测试, 评估所提方法在不同任务条件下的性能表现, 并进一步探讨模型中各个关键组件对整体性能的贡献.

3.1 数据集介绍

为支持人-物一体化检测算法的训练与评估, 本文采用两类典型的城市边缘视频检测数据集: EBPersons[18] 与 FADE[19], 分别面向“建筑边缘人员检测”与“高空坠落物检测”任务. 两者均采集自真实或模拟的高层建筑监控视角, 涵盖多种天气、光照与场景条件, 具备良好的时序特征和应用挑战性.

EBPersons 数据集包含 1314 段视频、超 8 万帧图像, 主要聚焦阳台、天台等高空区域中人员临边行为 (如站立、探身、攀爬等) 的检测任务, 具有目标尺度小、遮挡严重、姿态多变等典型特征. FADE 数据集共收集 1881 段高空坠物视频, 标注了瓶子、书本、工具、包装袋等 8 类典型坠落物体, 强调运动轨迹复杂、速度快、持续时间短等时序检测挑战.

为统一处理两个任务场景, 本文在不更改视频内容的前提下, 对 EBPersons 与 FADE 的标注结构进行统一整理, 采用统一的帧级边界框标注格式, 并整合为包含“人+物”目标类别的联合检测任务数据集. 该处理方式确保了数据集在类别命名、标注维度与评估流程上的一致性, 为构建统一建模框架提供了稳定基础.

3.2 评估指标

本文采用目标检测领域常用的平均精度均值 (mean average precision, *mAP*) 作为评价指标, 用以衡量模型在多类别、多检测阈值下的整体性能表现. 在单类检测任务中, 首先根据不同置信度下的检测结果绘制 Precision-Recall 曲线, 计算其面积作为该类的平均精度 (*AP*):

$$AP_c = \int_0^1 P_c(r)\mathrm{d}r \tag{6}$$

对于 C 类检测任务, *mAP* 定义为所有类别 *AP* 的平均值:

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{7}$$

此外, 真实框与预测框之间的重叠程度通过 *IoU* (intersection over union) 定义:

$$IoU = \frac{A_{\mathrm{pred}} \cap A_{\mathrm{gt}}}{A_{\mathrm{pred}} \cup A_{\mathrm{gt}}} \tag{8}$$

当 $IoU \geq$某一阈值 (本文统一设定为 0.5) 时, 预测框被视为正确命中 (true positive).

3.3 多任务性能检测对比

3.3.1 训练设置与模型初始化

本文所提出的检测框架基于 Faster R-CNN 主干网络, 并引入运动感知模块 (MAM)、时间区域兴趣点对齐操作符 (TROI Align) 与序列级语义聚合头部 (SELSA Head) 这 3 个模块. 主干网络采用在 ImageNet 上预训练的 ResNet-50 模型作为初始化; 其余模块 (如 SELSA Head) 采用 Xavier 初始化方式. 整个网络在本文构建的数据集上进行端到端微调训练, 分别在人员检测子任务与坠物检测子任务上进行评估.

3.3.2 损失函数设计

考虑到高空临边人员和小型坠物目标都属于小目标, 且在视频序列中存在类似的检测难点, 为简化训练并保证两类目标的平衡检测, 本文对人员和坠物采用统一的损失函数. 具体为:

$$L = \frac{1}{N}\sum_{i=1}^{N} [L_{\mathrm{cls}}(p_i, p_i^*) + \lambda L_{\mathrm{reg}}(t_i, t_i^*)] \tag{9}$$

其中, p_i与p_i^*为第i个 ROI 的预测类别与真实标签, t_i与t_i^*为预测边框与真实边框, $L_{\mathrm{cls}}(\cdot)$表示分类部分的交叉熵损失函数, $L_{\mathrm{reg}}(\cdot)$表示回归部分的平滑L_1损失函数, λ为权重系数.

3.3.3 多任务性能对比结果

为进一步评估本文所提出的人-物一体化检测框

架在边缘安全场景中的性能优势, 本文选取了多种主流目标检测方法和视频目标检测方法进行对比实验, 包括通用图像检测模型、行人检测模型以及具备时间建模能力的视频目标检测模型. 所有模型均在本文构建的数据集上进行训练与评估, 采用统一的实验设置与评估指标, 确保结果的可比性与公平性.

具体选取的对比方法包括: Faster R-CNN[5]、YOLOX[22]、DETR[23]等通用目标检测算法及 FGFA[10]、SELSA[21]、TROI[12]等典型视频目标检测方法. 其中, 视频类模型均采用与本文相同的主干网络 (ResNet-50), 并在多帧输入条件下运行, 确保检测场景一致. 实验结果如表 1 所示. 可以看出, 传统静态图像检测模型 (如 Faster R-CNN 和 YOLOX) 在边缘场景中表现相对较弱, 平均精度较低. 引入时序建模能力的视频检测方法 (如 FGFA、SELSA 和 TROI) 整体性能有所提升, 但仍存在融合策略局限或语义不稳定的问题. 相比之下, 本文提出的融合 MAM、TROI Align 与 SELSA Head 的统一检测框架在人员检测和坠物检测两个任务中均取得最优性能, 验证了所提方法在复杂边缘视频场景下的通用性与稳定性.

表 1 不同检测方法在人-物一体化任务中 *mAP*@50 实验结果 (%)

模块组合	人员检测 *mAP*@50	物体检测 *mAP*@50	人-物 *mAP*@50
Faster R-CNN+FPN[5]	53.5	18.3	35.2
YOLOX[22]	53.2	15.5	33.4
DETR[23]	54.2	14.3	32.0
TROI[12]	58.2	4.5	26.5
SELSA[21]	57.4	5.6	27.2
FGFA[10]	57.0	3.4	28.5
MOG2[20]	57.6	10.6	32.8
本文方法	58.5	39.2	55.4

3.4 消融实验

为全面评估各时间建模模块在人员检测与坠物检测任务中的作用, 本文设计了针对 3 个组件的不同网络结构进行消融对比实验, 实验结果如表 2 所示. 这 3 个模块均在不同维度提升了检测性能, 经分析: MAM 显著增强了模型对短时、高速运动目标 (如坠落物) 的响应能力; TROI Align 有效缓解了因视角偏移与遮挡引起的空间特征漂移; SELSA Head 通过语义对齐稳定了跨帧特征表达. 值得注意的是, 这 3 个模块并非孤立工作, 而是在不同维度形成互补: MAM 侧重于短时运动显著性增强, 使坠落物等快速目标能够被有效捕捉; TROI Align 在空间层面保证特征跨帧对齐, 缓解因视角变化和遮挡导致的特征漂移; SELSA Head 则在语义层面进一步聚合上下文信息, 提升跨帧特征的一致性. 三者协同作用下, 模型能够同时兼顾短时动态建模、空间稳定性与语义一致性, 从而在复杂场景中保持鲁棒的检测性能. 最终模型融合三者, 在小目标、遮挡和运动模糊等复杂条件下均表现出最优性能.

表 2 消融实验结果

消融组合	运动感知模块MAM	时间区域兴趣点对齐操作符 TROI Align	序列级语义聚合头部 SELSA Head	人-物 *mAP*@50 (%)
1	×	×	×	49.5
2	√	×	×	51.1
3	×	√	×	53.5
4	×	×	√	52.6
5	√	√	×	54.7
6	√	×	√	54.1
7	×	√	√	54.5
8	√	√	√	55.4

值得一提的是, 尽管 3 类模块分别源于不同检测任务, 但其功能在统一网络结构下表现出良好的互补性, 验证了所提出人-物一体化检测框架在“统一结构、共享模块”的设计理念下具有良好的泛化能力与工程适配潜力.

3.5 实时性及效率分析

为验证所提方法在实际部署中的性能, 本文将其与 Faster R-CNN 基线进行了对比, 结果如表 3 所示. 在 NVIDIA RTX 4090 平台上, Faster R-CNN 的推理速度为 9.3 f/s, 而本文方法由于引入时间建模模块, 推理速度略降至 8.7 f/s. 但在精度方面, 本文方法显著优于 Faster R-CNN, 能够更好地应对小目标检测与复杂场景识别问题. 整体而言, 本文方法在保持较高检测精度的同时, 仍具备接近实时的处理效率, 展现出良好的应用价值与实际部署潜力.

表 3 不同检测方法实时性对比分析

模型	推理速度 (f/s)
Faster R-CNN[5]	9.3
本文方法	8.7

4 结论与展望

本文针对智慧城市建筑边缘场景下频发的人员临边与高空坠物事件, 提出了一种统一的人-物一体化边缘安全检测框架. 该方法以 Faster R-CNN 为基础, 集成运动感知模块 (MAM)、时间区域兴趣点对齐操作符

(TROI Align) 及序列级语义聚合头部 (SELSA Head), 从运动显著性、空间位置对齐到语义建模引导这 3 方面提升模型的时序感知能力与检测鲁棒性.

在实验部分, 本文基于 EBPersons 与 FADE 数据集统一构建了一个人-物联合标注的视频检测方法, 完成了多任务检测性能验证与关键模块消融分析. 实验结果表明, 所提方法在人员检测与坠物检测两个子任务中均取得不错的性能, 体现出良好的通用性与跨任务适应能力.

尽管取得了一定的进展, 当前方法仍存在一些局限, 如在低照度、强遮挡或长时间连续视频中的鲁棒性仍需提升, 且实际部署时在推理效率与资源消耗方面仍有优化空间. 未来工作中, 可进一步引入多模态信息 (如深度图、热成像)、行为理解与目标追踪机制, 实现更高效、更智能的边缘风险事件综合感知, 为智慧城市的安全监控系统提供更加全面的技术支撑.

参考文献

1 房毓菲, 单志广. 智慧城市顶层设计方法研究及启示. 电子政务, 2017(2): 75–85.

2 黄凯奇, 陈晓棠, 康运锋, 等. 智能视频监控技术综述. 计算机学报, 2015, 38(6): 1093–1118.

3 张艳, 张明路, 吕晓玲, 等. 深度学习小目标检测算法研究综述. 计算机工程与应用, 2022, 58(15): 1–17.

4 何平, 李刚, 李慧斌. 基于深度学习的视频异常检测方法综述. 计算机工程与科学, 2022, 44(9): 1620–1629.

5 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

6 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

7 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2999–3007.

8 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.

9 Everingham M, Van Gool L, Williams CKI, *et al*. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

10 Zhu XZ, Wang YJ, Dai JF, *et al*. Flow-guided feature aggregation for video object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 408–417.

11 Chen YH, Cao Y, Hu H, *et al*. Memory enhanced global-local aggregation for video object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10334–10343.

12 Gong T, Chen K, Wang XJ, *et al*. Temporal ROI align for video object recognition. Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 1442–1450.

13 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.

14 Zhang SS, Benenson R, Schiele B. CityPersons: A diverse dataset for pedestrian detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4457–4465.

15 Shao S, Zhao ZJ, Li BX, *et al*. CrowdHuman: A benchmark for detecting human in a crowd. arXiv:1805.00123, 2018.

16 Dollár P, Wojek C, Schiele B, *et al*. Pedestrian detection: A benchmark. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 304–311.

17 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3354–3361.

18 EBPersons. https://ebpersons.github.io/index.html. [2025-09-05].

19 Tu ZG, Zhang ZB, Gao ZT, *et al*. FADE: A dataset for detecting falling objects around buildings in video. IEEE Transactions on Information Forensics and Security, 2025, 20: 9746–9759. [doi: 10.1109/TIFS.2025.3607254]

20 Zivkovic Z. Improved adaptive Gaussian mixture model for background subtraction. Proceedings of the 17th International Conference on Pattern Recognition. Cambridge: IEEE, 2004. 28–31.

21 Wu HP, Chen YT, Wang NY, *et al*. Sequence level semantics aggregation for video object detection. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9216–9224.

22 Ge Z, Liu ST, Wang F, *et al*. YOLOX: Exceeding YOLO series in 2021. arXiv:2107.08430, 2021.

23 Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with Transformers. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229.

(校对责编: 李慧鑫)